DOCUMENT RESUME

ED 343 949 TM 018 135

AUTHOR Livingston, Samuel A.
TITLE Small-Sample Equating with Log-Linear Smoothing.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-4
PUB DATE Jan 92
NOTE 40p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Advanced Placement Programs; College Entrance Examinations; *Equated Scores; High Schools; High School Students; Mathematical Models; *Sample Size; Sampling; Student Placement; *Test Construction
IDENTIFIERS Advanced Placement Examinations (CEEB); *Chained Equipercentile Equating; Log Linear Models; *Smoothing Methods

ABSTRACT

This study investigated the extent to which log-linear smoothing could improve the accuracy of common-item equating by the chained equipercentile method in small samples of examinees. Examinee response data from a 100-item test (the Advanced Placement Examination in United States History) were used to create two overlapping forms of 58 items each, with 24 items in common. The criterion equating was a direct equipercentile equating of the two forms in the full population of 93,283 high school students. Anchor equatings were performed in samples of 25, 50, 100, and 200 examinees, with 50 pairs of samples at each size level. Four equatings were performed with each pair of samples: one based on unsmoothed distributions and three based on varying degrees of smoothing. Smoothing reduced, by at least half, the sample size required for a given degree of accuracy. Smoothing that preserved only two moments of the marginal distributions resulted in equatings that failed to capture the curvilinearity in the population equating. A list of nine references, two tables, 16 figures, and an appendix giving a formula used in the analysis are included. (Author/SLD)

RESEARCH REPORT



# SMALL-SAMPLE EQUATING WITH LOG-LINEAR SMOOTHING

**Samuel A. Livingston**

**Educational Testing Service**
**Princeton, New Jersey**
**January 1992**

# Small-Sample Equating with Log-Linear Smoothing

Samuel A. Livingston

Educational Testing Service

ACKNOWLEDGMENTS

I thank Kirsten Yocom for creating and running the computer programs that made this study possible. I also thank Gerald Melican and Marna Golub-Smith for their help in planning the study, Helen Kahn and Carole Bleistein for their help in obtaining and preparing the data, and Ted Blew for producing the computer-drawn graphs. Support for this research was provided by the Office of Corporate Quality Assurance at Educational Testing Service.

ABSTRACT

This study investigated the extent to which log-linear smoothing could improve the accuracy of common-item equating by the chained equipercentile method in small samples of examinees. Examinee response data from a 100-item test were used to create two overlapping forms of 58 items each, with 24 items in common. The criterion equating was a direct equipercentile equating of the two forms in the full population of 93,283 examinees. Anchor equatings were performed in samples of 25, 50, 100, and 200 examinees, with fifty pairs of samples at each size level. Four equatings were performed with each pair of samples: one based on unsmoothed distributions and three based on varying degrees of smoothing. Smoothing reduced, by at least half, the sample size required for a given degree of accuracy. Smoothing that preserved only two moments of the marginal distributions resulted in equatings that failed to capture the curvilinearity in the population equating.

Small-Sample Equating with Log-Linear Smoothing

## The problem

Like other statistical operations, the equating of test scores is subject to sampling error. When we equate test forms, the test-takers whose scores serve as the raw material for our calculations are, in most cases, only a sample of the population of test-takers for whom we want the equating to be correct. If the sample is large and representative of the population, the equating relationship in the sample is likely to be a precise estimate of the equating relationship in the population. But if the sample is small, it may be a very imprecise estimate. If the equating is based on an anchor design, with a different sample of test-takers taking each of the forms to be equated, the imprecision is compounded.

This problem -- the instability of anchor equating results in small samples of examinees -- is a special case of the more general problem of small-sample instability. A common approach to the problem is to use a strong model, with only a small number of parameters to be estimated from the data. Linear equating methods can be considered an example of this approach. For example, the Tucker method (Angoff, 1984, pp. 110-111; Petersen, et al, 1989) uses only the means, standard deviations, and intercorrelation of the scores on the test to be equated and the anchor test, in each sample of test-takers. The disadvantage of linear methods is their lack of flexibility; they cannot reproduce a curvilinear relationship between the tests to be equated. When the equating relationship in the population is curvilinear, the results of a linear equating may be highly inaccurate in some regions of the score range, usually at the extremes.

Unfortunately, curvilinear equating relationships are not rare. The same factors that make equating necessary -- a difference in the difficulty of test forms -- tend to produce curvilinear equating relationships. Typically, the harder form tends to spread out the scores of the stronger test-takers, while the easier form tends to spread out the scores of the weaker test takers. As a result, the slope of the equating relationship between the score distributions on the two forms tends to change from the lower portion of the score range to the higher portion.

## The approach

Fortunately, linear equating is not the only way to use a strong model in equating test scores. Another way is to use a strong model to estimate the score distributions in the population from the distributions observed in the samples. The estimated population distributions can then be used in place of the observed sample distributions in computing the equating relationship. The score distributions estimated for the population are typically much "smoother" than those observed in the sample (i.e., the frequencies change more gradually from one score level to the next). Therefore, the estimated distributions are often described as "smoothed", and the process is often referred to as "smoothing".

Some smoothing models offer the user a choice in the number of parameters to be estimated from the data. Log-linear models (Holland and Thayer, 1987) offer this kind of flexibility. The user specifies the moments of the observed distributions to be preserved in the smoothed distributions. The algorithm then computes the smoothest distribution that has those specified moments.[1] These models can be as strong as they need to be. If the samples are very small, the smoothed distribution may preserve only the mean and standard deviation of the observed distribution. If the samples are fairly large, a larger number of moments of the observed distribution can be preserved. In all cases, the smoothed distribution is a discrete distribution over the same range of possible scores as the observed distribution.

This paper describes a study intended to determine the extent to which the use of log-linear models for estimating score distributions would reduce the number of test-takers required for equating test scores with reasonable accuracy. The method of equating was the chained equipercentile method. This method is the composition of two separate equipercentile equatings; each form of the test is equated to the anchor test in the sample of test-takers taking that form. Figure 1 describes the method graphically. This method will produce accurate results if the equating relationship between each form and the anchor is the same in the sample of test-takers taking that form as in the population (i.e., the population of test-takers for which the equating relationship is to be determined). Livingston, Dorans, and Wright (1990) found empirically that this method (unlike some others) tended to be free from bias when one of the samples was not representative of the population. The disadvantage of the chained equipercentile method is its sensitivity to sampling variability. To overcome this disadvantage, it is necessary to replace the score distributions observed in the samples of test-takers with better estimates of the score distributions in the population. Log-linear models may offer a solution to this problem.

## Previous studies

Of the studies that have been done, the one most directly relevant to this investigation is by Hanson (1991). That study investigated the accuracy of equating by a somewhat different equipercentile method, called "frequency estimation", in which the anchor test is used as a stratifying variable in estimating the score distributions to be equated. Hanson used two log-linear smoothing models, as well as two other strong models (beta-binomial and beta-compound-binomial), to smooth the score distributions before equating by the frequency estimation equipercentile method. Hanson also equated by three linear equating methods. He compared the results of each procedure, in samples ranging in size from 100 to 3000, with the results of a criterion equating that used the data from all available test-takers -- about 39,000 for each form. (The criterion equating also used a log-linear model to smooth the score distributions before equating.) All four smoothing methods tended to improve the accuracy of the frequency estimation equipercentile equating. The smoothing method that produced the best results was a log-linear smoothing

---

[1]See Rosenbaum and Thayer (1987, pp. 45-46) for a more precise statement of the criterion by which this estimated distribution is smoothest.

that preserved only the means, variances, skewnesses, and correlation of the observed bivariate distributions. However, in the smaller samples, one of the linear methods (the Tucker method) produced more accurate results than any version of the frequency estimation equipercentile equating, despite some curvature in the criterion equating functions.

Livingston and Feryok (1987) also investigated the effect of log-linear smoothing on equipercentile equating by frequency estimation, but with only three replications at each sample size level. With samples of 100 and 300 test-takers, the log-linear smoothing substantially improved the accuracy of the equating, i.e., its agreement with the results of equating in the full population. With samples of 1000 and 3000 test-takers, it did not.

Other studies on the use of smoothing methods in equating are less relevant to the investigation described in this report. Fairbank (1987) used a random-groups equating design, with no anchor. His smoothing methods included smoothing by medians, smoothing by weighted moving averages, and one strong model (negative hypergeometric distribution). Kolen and Jarjoura (1987) investigated the smoothing of the equating transformation itself, rather than the smoothing of the distributions prior to equating.

## The present study: method

The present study was designed to create a situation in which the equating relationship in the population was known. The data were taken from the responses of 93,283 high school students to the multiple-choice section of the Advanced Placement Examination in United States History. From the 100 items in this section of the examination, we created two overlapping sub-forms of 58 items each, with 24 items appearing in both sub-forms to serve as an anchor for equating. The test forms were constructed to be as similar in content as possible, while differing systematically in difficulty. The content specifications for this examination classify the items into three historical periods (before 1789, 1789 to 1914, 1915 to the present) and four historical categories (social/economic, political, diplomatic, cultural/intellectual). In each of the twelve possible cross-classifications, we specified one-fourth of the number of items available as the number to be common to both sub-forms. The remaining items were then divided equally between the two sub-forms. Where the number of remaining items was not divisible by two, one item was left out. Therefore, in each possible cross-classification the two sub-forms contained exactly the same number of items, and the common-item anchor test reflected that content distribution as closely as possible. For each of the 93,283 test takers, we computed three scores: a score on each of the two 58-item sub-forms and a score on the 24-item anchor test. The score was simply the number of items answered correctly.[2]

---

[2]Although the Advanced Placement Examinations actually are scored with a correction for guessing, we chose not to use the correction for guessing in our study, because it introduces additional complications into the definition of the equating transformation.

The attempt to create forms that differed in difficulty was successful. We arbitrarily labeled the more difficult of the two sub-forms "Form A" and the less difficult sub-form "Form B". The mean raw scores in the full population of test-takers were 29.2 items correct (50%) on Form A and 35.9 (62%) on Form B. The standard deviations were 8.7 items on Form A and 8.8 on Form B. Thus, the difference in the mean scores was about three-fourths of a standard deviation. None of the 93,283 test-takers achieved the maximum possible score of 58 items correct on Form A, and only 5 did so on Form B.

Also arbitrarily, we structured the study so that Form A was to be equated to Form B. That is, we sought a transformation that would determine, for each integer score on Form A, the corresponding score (not necessarily an integer) on Form B. To find the equating relationship of the two sub-forms in the population, we performed a direct equipercentile equating of the score distributions on Forms A and B in the entire population of 93,283 test-takers. The anchor test played no part in the determination of the population equating. The results of this direct, full-population equating served as the criterion for evaluating the results of the equatings based on samples of examinees.

Figure 2 shows the equating relationship of Forms A and B in the full population. The relationship is clearly curvilinear, as might be expected from the difference in difficulty. At the lower ability levels, Form B, the easier form, provides finer discrimination. For example, the difference between the 5th and 20th percentiles (of the same population) is only 6 points on the raw score scale for the more difficult Form A (the horizontal scale in Figure 2) but 8 points on the scale for the easier Form B (the vertical scale). Therefore, the slope of the equating curve that transforms the Form A scale to the Form B scale is greater than 1 in the lower portion of the range. At the higher ability levels, Form A, the harder form, provides finer discrimination, and the slope of the equating curve is less than 1.

We evaluated the smoothing methods by the following procedure:

1. Select a pair of samples of a specified size: 25, 50, 100, or 200 test-takers.

2. Arbitrarily associate each sample with one of the two sub-forms. For test-takers in the "Form A" sample, treat the score on Form B as unknown; for test-takers in the "Form B" sample, treat the score on Form A as unknown.

3. For each sample, smooth the joint distribution of scores on the sub-form and the anchor test. Perform three smoothings of each distribution. In the first smoothing, preserve only the first two univariate moments (means and standard deviations) and the first bivariate moment (the correlation between the sub-form and the anchor test). In the second smoothing, preserve these moments and also the third univariate moments (skewness). In the third smoothing, preserve also the fourth univariate moment (kurtosis). We will refer to these smoothings as the "two-moment" smoothing, the "three-moment" smoothing, and the "four-moment" smoothing.

4. In each pair of samples, perform a chained equipercentile equating of Form A to Form B through the anchor test, by equating Form A to the anchor test in the "Form A" sample and equating the anchor test to Form B in the "Form B" sample. Use the unsmoothed distributions in these equatings.

5. Repeat Step 4, but use the distributions created by the four-moment smoothing.

6. Repeat Step 5, but use the distributions created by the three-moment smoothing.

7. Repeat Step 6, but use the distributions created by the two-moment smoothing.

This procedure was replicated 50 times at each of the four specified sample size levels.

The results

A statistic that provides a good indication of the overall accuracy of an estimation procedure is the root-mean-squared deviation (RMSD) of the estimates from the quantity to be estimated. In this case, the quantity to be estimated is the equated score on Form B that corresponds to a given raw (number-correct) score on Form A, as determined by the direct equipercentile equating in the full population of test-takers. Figure 3a shows the accuracy of the equating in the samples of 200 test-takers. The accuracy statistic is the conditional RMSD, conditioned on the score to be equated and computed over the 50 replications of the equating experiment. (See Formula 1 in the Appendix.) The conditional RMSD is shown as a function of the raw score on Form A, for each of the four degrees of smoothing. The conditional RMSD is expressed in terms of points on the raw-score scale of Form B, the form to which Form A is being equated. One point is about one-tenth of the population standard deviation.

In Figure 3a, notice that the line representing the equating based on unsmoothed distributions does not extend to the ends of the score range. The reason is that in many of the samples, the unsmoothed distributions did not extend over the full range of scores, and the equipercentile equating relationship was undefined at the ends of the score range. The conditional RMSD is shown on the graph only for those score levels for which the conversion was defined in at least 45 of the 50 replications of the equating. This problem did not occur with the equating of the smoothed distributions, because the smoothed distributions extended over the full score range even when the unsmoothed distributions did not. Consequently, the equating of these distributions was defined over the full score range, and the conditional RMSD could be computed for all scores observed in the population.

Figure 3a shows that even with samples of 200 examinees, both the three-moment and four-moment smoothings provide a substantial improvement over the equating of unsmoothed distributions, as indicated by the smaller conditional RMSD of the sample equating from the population equating. The three-moment smoothing produced more accurate results than the four-moment smoothing in the

very high and very low portions of the score range, where the data tend to be sparse, and performed about as well as the four-moment smoothing in the rest of the score range. The equating based on the two-moment smoothing appears to have been about as accurate as the equating based on unsmoothed distributions.

Figure 3b shows the results for the samples of 100 examinees. All three smoothings appear to offer some degree of improvement over the equating of unsmoothed distributions; the three-moment and four-moment smoothings offer a substantial improvement throughout the score range. The three-moment smoothing appears to perform slightly better than the four-moment smoothing, particularly at the extremes of the score range.

Figure 3c shows the results for the samples of 50 examinees. All three smoothings produced a substantial improvement over the equating of unsmoothed distributions. The improvement produced by the smoothing was greater with samples of 50 examinees than with the larger samples, as might be expected. Again, the three-moment smoothing appears to produce the best results. Throughout most of the score range, it reduced the conditional RMSD by about one-third, as compared with the equating of unsmoothed distributions.

Figure 3d shows the results for the samples of 25 examinees. The improvement produced by smoothing was greater still, for all the smoothings. The two-moment and three-moment smoothings appear to work about equally well -- and somewhat better than the four-moment smoothing -- throughout most of the score range. The three-moment smoothing performed slightly better than the two-moment smoothing in the middle of the range; the two-moment smoothing performed better at the high end of the range.

Figures 3e, 3f, and 3g each compare two of the conditional RMSD curves from Figures 3a-3d. Figure 3e shows the conditional RMSDs for the equating in samples of 200 examinees without smoothing and in samples of 100 examinees with the three-moment smoothing. Figure 3f shows the same comparison for the samples of 100 examinees without smoothing and samples of 50 examinees with the three-moment smoothing. Figure 3g shows the same comparison for the samples of 50 examinees without smoothing and samples of 25 examinees with the three-moment smoothing. These comparisons show that the three-moment smoothing produces a greater increase in accuracy than could be achieved by doubling the size of the samples of examinees.

The results described above were obtained by conditioning on the score level (on the test to be equated) and computing the conditional RMSD over the fifty replications of the equating procedure. Another way to look at the data is to compute an RMSD over all score levels <u>for each replication</u> of the equating. This overall RMSD is computed over the population of test-takers. (See Formula 2 in the Appendix.) Figure 4 shows the distributions, over the fifty replications, of this RMSD, for each combination of sample size and smoothing method. The means and standard deviations of these distributions are shown in Table 1. Notice that the overall RMSD is not shown for the equatings of unsmoothed distributions. To compute the overall RMSD, it is necessary that the equated score be defined for all the test-takers in the population. In general, the equating based on the unsmoothed sample score

distributions did not produce equated scores for the test-takers with the highest and lowest scores. Therefore, the overall RMSD could not be computed.

Figure 4 shows the tendency of the RMSD to decrease as the sample size increases. It also shows the extent to which the accuracy of the equating tended to vary as a result of the sampling of examinees, and how this variation decreased as the sample size increased. However, our main interest is in the comparison of the three smoothings. With samples of 25 examinees, the two-moment and three-moment smoothings appear to have produced somewhat more accurate results than the four-moment smoothing. With samples of 50 and 100 examinees, the three-moment smoothing produced the most accurate results. With samples of 200 examinees, the three-moment smoothing still produced the best results, although the four-moment smoothing performed nearly as well. These conclusions from the graphs in Figure 4 are corroborated by the means shown in Table 1.

It is possible to look at the results in more detail, by conditioning on a single score on Form A and looking at the results of the individual replications of the equating procedure. Each replication produced an equated score which can be considered an estimate of the equated score produced by the direct population equating. If the anchor equating in the samples of examinees were exactly correct at this score level, it would produce the same result as the direct equating in the population; the difference between the results of the two equatings (sample and population) would be zero. If the equating procedure is highly accurate at the specified score level, these differences will be tightly clustered around zero.

Figure 5a shows distributions of the difference between the sample and population equatings in determining the score on Form B that corresponds to a score of 20 on Form A. Each of these conditional distributions contains fifty data points, one for each replication of the equating. There is a separate distribution for each combination of sample size and smoothing method. The distributions in Figure 5a are conditional distributions because they apply only to score level 20 on Form A. Figures 5b-5e show similar distributions for other score levels: 25, 30, 35, and 40. Table 2 shows the mean and standard deviation of each of the conditional distributions in Figures 5a-5e. The mean of the conditional distribution can be interpreted as the empirically determined bias of the equating procedure at that score level. The standard deviation indicates the sampling variability of the equated score.

The most obvious (and least surprising) feature of the distributions in Figures 5a-5e is the decrease in the spread of the distribution as the sample size increases. The distributions also show the extent to which smoothing tended to improve the accuracy of the equating, offering a greater improvement when the samples were smaller.

A more interesting feature of these distributions concerns the two-moment smoothing; it appears to introduce a consistent bias into the equating. The equatings based on the two-moment smoothing tended to produce equated scores that were too low at scores 25 and 30 (Figures 5b and 5c) and too high at score 40 (Figure 5e). This is the same kind of bias that would have been

introduced by a linear equating method (which also uses only two moments of each distribution to determine the equating relationship). Preserving the third moment of the distributions greatly reduced this bias, at very little cost in terms of increased variation over replications, as can be seen in Table 2. Preserving the fourth moment tended to eliminate the bias, but the sampling variability of the equating results in the smaller samples (25 and 50) increased substantially at the lower score levels (20, 25, and 30).

Figure 6 provides a clearer illustration of the bias introduced by the smoothings that preserve fewer than four moments of the unsmoothed distributions. This graph shows the bias, at each score level, in the equatings based on samples of 200 examinees. Note the large bias in the results of the two-moment smoothing and the much smaller bias, in the same direction, in the results of the three-moment smoothing. In contrast, notice how closely the results of the four-moment smoothing follow those of the equatings based on the unsmoothed distributions, with very little bias.

Discussion

The smaller the sample, the more it is likely to differ from the population. Therefore, the benefits of smoothing are greatest when the sample is small. For the same reason, smaller samples call for smoothing methods that preserve fewer characteristics of the observed data. For smoothing test score distributions by the log-linear method, the issue is how many moments of the observed distribution to preserve in the smoothed distribution. To preserve too few moments is to lose some of the information contained in the sample. But if the sample moments are far enough from the corresponding population moments, preserving them may do more harm than good.

A statistician might claim that samples as small as 25 are too small for accurate estimation of the third moment (skewness) of the population distribution. However, in this application of log-linear smoothing, preserving the third moment of the observed distributions improved the accuracy of the equating results, even in the samples of 25 examinees.

It might then seem that with samples of 200 examinees, preserving the fourth moment of the observed distributions would improve the accuracy of the equating (in comparison with preserving only three moments). In this study, preserving the fourth moment in samples of 200 examinees did not improve the overall accuracy of the equating, but it did remove a small bias that was present in the equating results based on three-moment smoothing. Systematic bias can be more harmful than the same amount of random error when equating results are linked to maintain a score scale through several forms of a test. The danger is that if the difficulty of the test increases (or decreases) steadily over several test forms, the bias in the equating will be in the same direction each time and will tend to accumulate. The result will be "scale drift".

One feature of equatings based on distributions produced by log-linear smoothing is that, like linear equatings, they span the full range of possible scores, extending beyond the highest and lowest scores observed in the samples. But how accurate are the equatings in these regions beyond the

data? Figures 3a to 3d suggest that, for any given sample size, the equatings based on the three-moment smoothings are at least as accurate at the high end of the distribution, where data are sparse or absent, as the equatings based on unsmoothed distributions are in the middle region, where there is plenty of data. At the extreme low end of the score range, the RMSD becomes fairly large even for the smoothed distributions, especially when the sample size is small. However, in this region, the equating based on unsmoothed distributions cannot be computed in most of the samples. A user who needed to determine an equating in this range would have some alternatives: extrapolate from the equating based on unsmoothed distributions, use a linear equating, or assign converted scores arbitrarily. None of these alternatives seems preferable to the use of the equipercentile equating based on smoothed distributions.

Do the results of this study justify the equating of test forms by the chained equipercentile method in samples as small as 50 or even 25? The overall RMSD of the equating based on samples of 25 examinees was about 0.15 standard deviations; for the samples of 50 examinees, it was about 0.10 standard deviations. However, the RMSD in the extremes of the score range was much larger than the overall RMSD. Whether these results are acceptable depends on a number of factors, including the range of scores over which accuracy is required, the availability of larger samples (and the cost of obtaining them), and the probable consequences of reporting scores that are not equated. The results of this study indicate that, whatever degree of accuracy is required, log-linear smoothing makes it possible to equate test forms with samples of about half as many examinees as would be required if no smoothing were done.

It seems appropriate to close with a word of caution. The present study involved one test, one population of examinees, and one equating method -- hardly a basis for any sweeping conclusions. It seems reasonable to expect that similar studies with different tests and examinee populations would tend to corroborate these results. Generalizing to another method of equating (e.g., by frequency estimation) would be more speculative, although it seems likely that the results of similar studies with a different equating method would follow a similar pattern. In the meantime, the results of this study can serve as a guide to those who must equate test scores on the basis of data from small samples of examinees.

## REFERENCES


Angoff, W. H. (1984) Scales, Norms, and Equivalent Scores. Princeton, NJ: Educational Testing Service.

Fairbank, B. A. (1987) The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. Applied Psychological Measurement, 11, 245-262.

Hanson, B. A. (1991) A comparison of bivariate smoothing methods in common-item equipercentile equating. Unpublished manuscript. Iowa City, IA: The American College Testing Program.

Holland, P. W. and Thayer, D. T. (1987) Notes on the use of log-linear models for fitting discrete probability distributions. Program Statistics Research Technical Report No. 87-79. Princeton, NJ: Educational Testing Service.

Kolen, M. J. and Jarjoura, D. (1987) Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. Psychometrika, 52, 43-59.

Livingston, S. A., Dorans, N. J., and Wright, N. K. (1990) What combination of sampling and equating methods works best? Applied Measurement in Education, 3, 73-95.

Livingston, S. A. and Feryok, N. J. (1987) Univariate vs. bivariate smoothing in frequency estimation equating. Research Report No. 87-36. Princeton, NJ: Educational Testing Service.

Petersen, N. S., Kolen, M. J., and Hoover, H. D. (1989) Scaling, norming, and equating. In Linn, R. L. (ed.), Educational Measurement, 3rd ed. New York: Macmillan.

Rosenbaum, P. R. and Thayer, D. (1987) Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. British Journal of Mathematical and Statistical Psychology, 40, 43-49.

Table 1.

Mean and standard deviation, over 50 replications, of the overall RMSD of sample equatings from the population equating.

| Sample size | | Smoothing | | |
|---|---|---|---|---|
| | | 4-moment | 3-moment | 2-moment |
| 25 | Mean | 1.69 | 1.47 | 1.51 |
| | S.D. | 0.49 | 0.48 | 0.37 |
| 50 | Mean | 1.28 | 1.05 | 1.24 |
| | S.D. | 0.58 | 0.54 | 0.36 |
| 100 | Mean | 0.89 | 0.80 | 1.07 |
| | S.D. | 0.29 | 0.28 | 0.19 |
| 200 | Mean | 0.58 | 0.56 | 0.98 |
| | S.D. | 0.29 | 0.28 | 0.15 |

Table 2.

Summary statistics for conditional distributions of differences between sample and population equatings.

| Score on Form A | Sample Size | Smoothing Method | Mean | Standard Deviation |
|---|---|---|---|---|
| 20 | 25 | Unsmoothed | -.22 | 3.03 |
| | | Four-moment | +.14 | 1.83 |
| | | Three-moment | +.35 | 1.44 |
| | | Two-moment | +.29 | 1.43 |
| | 50 | Unsmoothed | -.44 | 2.18 |
| | | Four-moment | -.35 | 1.53 |
| | | Three-moment | -.06 | 1.29 |
| | | Two-moment | -.10 | 1.20 |
| | 100 | Unsmoothed | -.12 | 1.40 |
| | | Four-moment | +.04 | 0.97 |
| | | Three-moment | +.15 | 0.82 |
| | | Two-moment | +.02 | 0.77 |
| | 200 | Unsmoothed | -.08 | 1.04 |
| | | Four-moment | -.15 | 0.71 |
| | | Three-moment | -.01 | 0.70 |
| | | Two-moment | -.09 | 0.65 |
| 25 | 25 | Unsmoothed | -.28 | 2.24 |
| | | Four-moment | +.10 | 1.45 |
| | | Three-moment | -.18 | 1.16 |
| | | Two-moment | -.47 | 1.16 |
| | 50 | Unsmoothed | +.04 | 1.65 |
| | | Four-moment | +.06 | 1.21 |
| | | Three-moment | -.27 | 0.91 |
| | | Two-moment | -.75 | 0.85 |
| | 100 | Unsmoothed | -.10 | 1.15 |
| | | Four-moment | +.02 | 0.83 |
| | | Three-moment | -.21 | 0.70 |
| | | Two-moment | -.68 | 0.60 |
| | 200 | Unsmoothed | -.08 | 0.71 |
| | | Four-moment | -.03 | 0.58 |
| | | Three-moment | -.28 | 0.54 |
| | | Two-moment | -.79 | 0.49 |

Table 2 (continued).

Summary statistics for conditional distributions of differences between sample and population equatings.

| Score on Form A | Sample Size | Smoothing Method | Mean | Standard Deviation |
|---|---|---|---|---|
| 30 | 25 | Unsmoothed | +.27 | 2.01 |
| | | Four-moment | +.16 | 1.28 |
| | | Three-moment | -.19 | 1.09 |
| | | Two-moment | -.47 | 1.10 |
| | 50 | Unsmoothed | +.32 | 1.14 |
| | | Four-moment | +.29 | 0.85 |
| | | Three-moment | -.14 | 0.70 |
| | | Two-moment | -.63 | 0.66 |
| | 100 | Unsmoothed | +.06 | 0.99 |
| | | Four-moment | +.05 | 0.69 |
| | | Three-moment | -.18 | 0.62 |
| | | Two-moment | -.62 | 0.55 |
| | 200 | Unsmoothed | +.03 | 0.65 |
| | | Four-moment | +.05 | 0.49 |
| | | Three-moment | -.22 | 0.43 |
| | | Two-moment | -.73 | 0.39 |
| 35 | 25 | Unsmoothed | -.08 | 1.90 |
| | | Four-moment | +.03 | 1.20 |
| | | Three-moment | +.06 | 1.17 |
| | | Two-moment | -.04 | 1.20 |
| | 50 | Unsmoothed | +.15 | 1.10 |
| | | Four-moment | +.16 | 0.74 |
| | | Three-moment | +.09 | 0.72 |
| | | Two-moment | -.07 | 0.76 |
| | 100 | Unsmoothed | -.01 | 0.84 |
| | | Four-moment | -.02 | 0.63 |
| | | Three-moment | -.03 | 0.60 |
| | | Two-moment | -.12 | 0.64 |
| | 200 | Unsmoothed | -.06 | 0.63 |
| | | Four-moment | -.03 | 0.41 |
| | | Three-moment | -.07 | 0.38 |
| | | Two-moment | -.21 | 0.39 |

Table 2 (continued).

Summary statistics for conditional distributions of differences between sample and population equatings.

| Score on Form A | Sample Size | Smoothing Method | Mean | Standard Deviation |
|---|---|---|---|---|
| 40 | 25 | Unsmoothed | +.04* | 2.28* |
| | | Four-moment | +.14 | 1.39 |
| | | Three-moment | +.49 | 1.47 |
| | | Two-moment | +.71 | 1.35 |
| | 50 | Unsmoothed | +.28 | 1.54 |
| | | Four-moment | +.07 | 1.03 |
| | | Three-moment | +.39 | 0.99 |
| | | Two-moment | +.81 | 0.99 |
| | 100 | Unsmoothed | -.01 | 0.96 |
| | | Four-moment | +.00 | 0.73 |
| | | Three-moment | +.25 | 0.71 |
| | | Two-moment | +.72 | 0.76 |
| | 200 | Unsmoothed | +.01 | 0.64 |
| | | Four-moment | -.06 | 0.38 |
| | | Three-moment | +.18 | 0.39 |
| | | Two-moment | +.65 | 0.45 |

*Based on only 48 replications; the equating transformation was undefined in two of the 50 replications.

## APPENDIX

### Formulas for the Root-Mean-Squared Deviation

Let j index the pairs of samples of a given size: j = 1, 2, ... , 50.

Let x represent a score on Form A.

Let $N_x$ represent the number of test-takers in the population with score x.

Let $y_x$ represent the score on Form B that equated to x in the direct equating in the population.

Let $\hat{y}_{xj}$ represent the score on Form B that equated to x in the anchor equating in the jth sample.

1. The conditional RMSD at score x (as in Figures 3a-3d) is computed by

$$RMSD(x) = \sqrt{\frac{1}{50}\sum_{j=1}^{50} (\hat{y}_{xj} - y_x)^2} \quad .$$

2. The overall RMSD for the jth replication (as in Figure 4) is computed by

$$RMSD(j) = \sqrt{\frac{\sum_x N_x (\hat{y}_{xj} - y_x)^2}{\sum_x N_x}} \quad .$$

**Figure 1. Illustration of the Chained Equipercentile Method of Equating.**

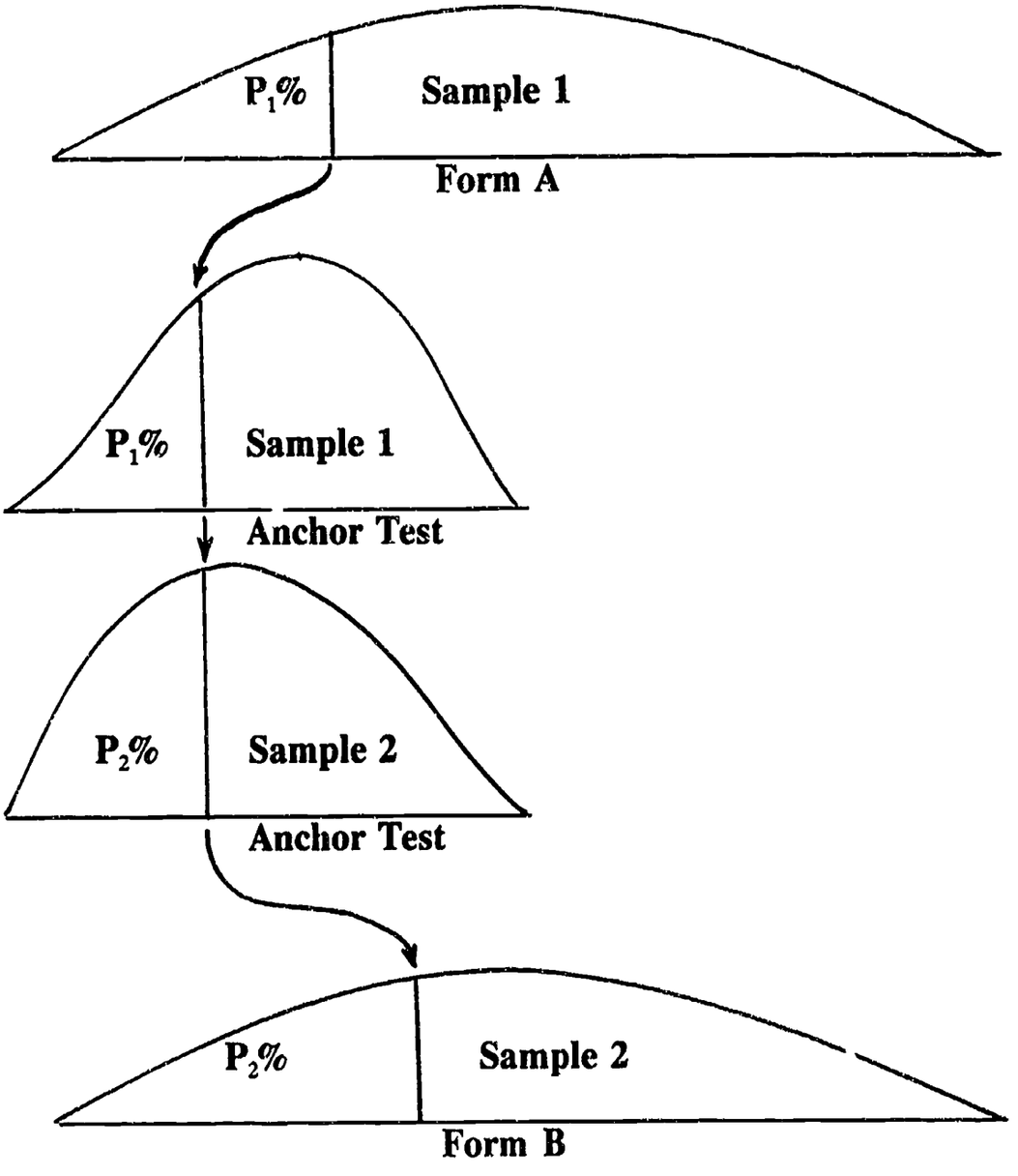


Figure 2.

**Results of direct equating of Form A to Form B in the full population.**

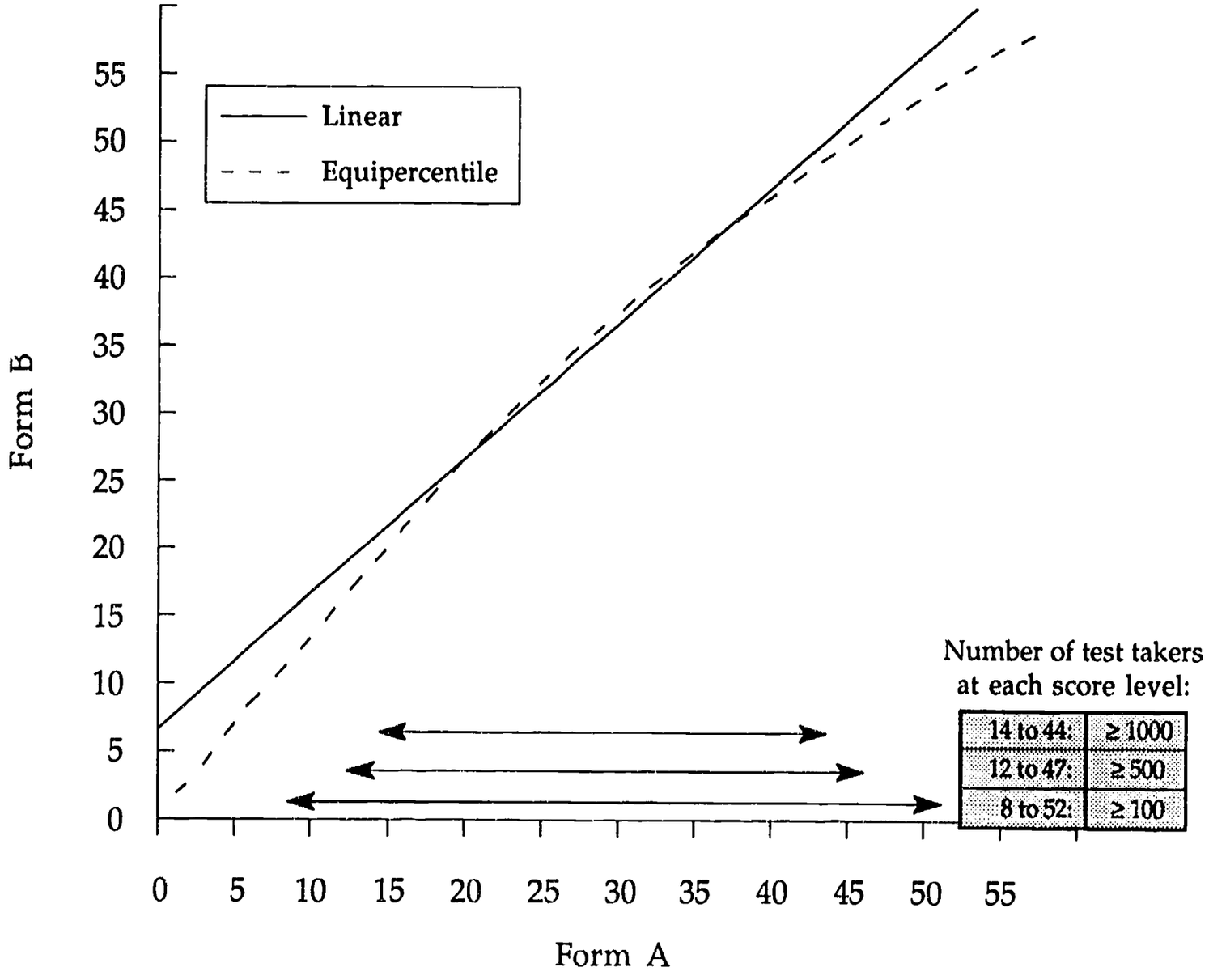


Figure 3. Conditional root-mean-square difference (RMSD) of sample equating (through common-item anchor) from population equating, over 50 replications of the equating procedure.

Figure 3a. Samples of 200 examinees.

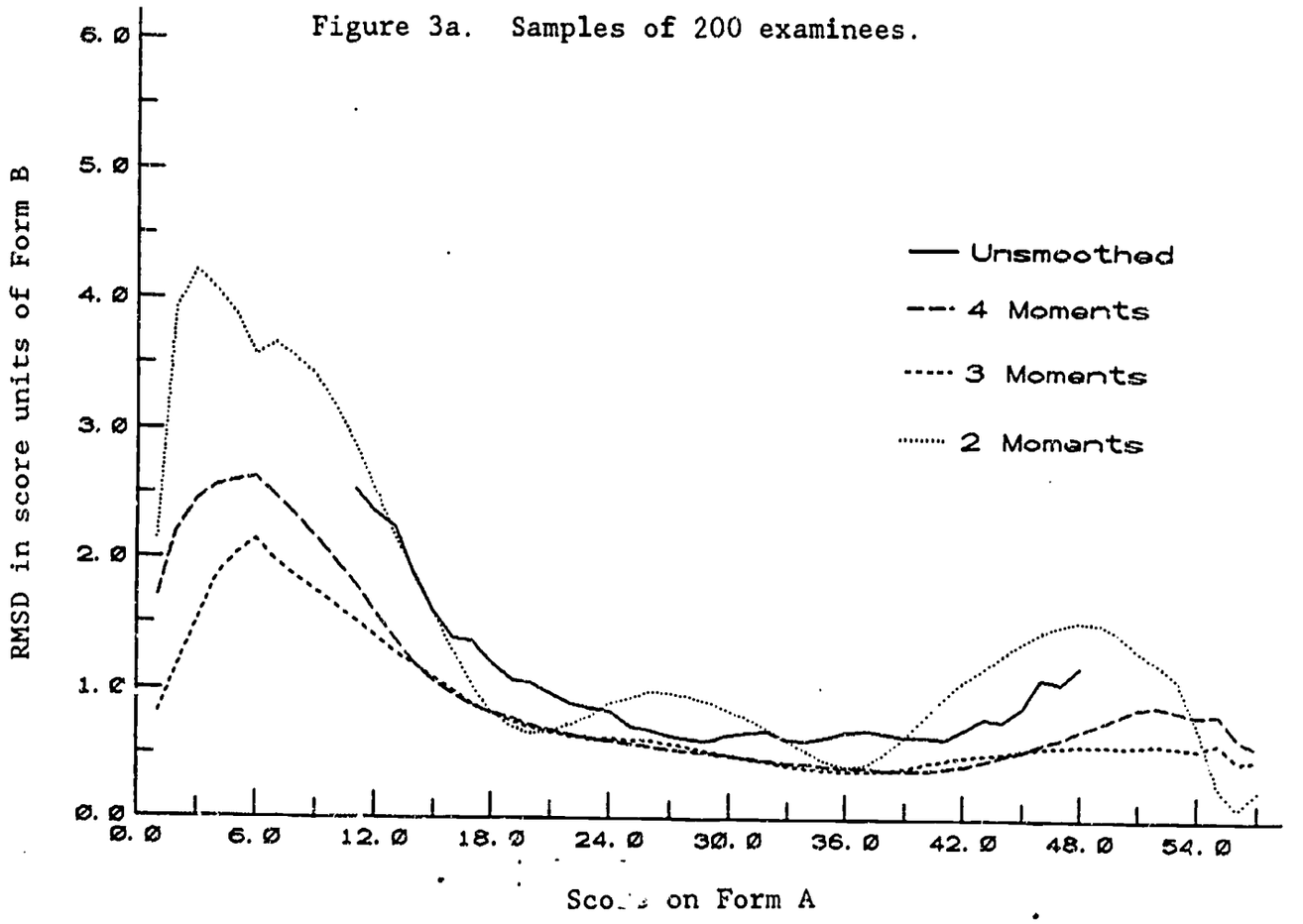



Figure 3b. Samples of 100 examinees.

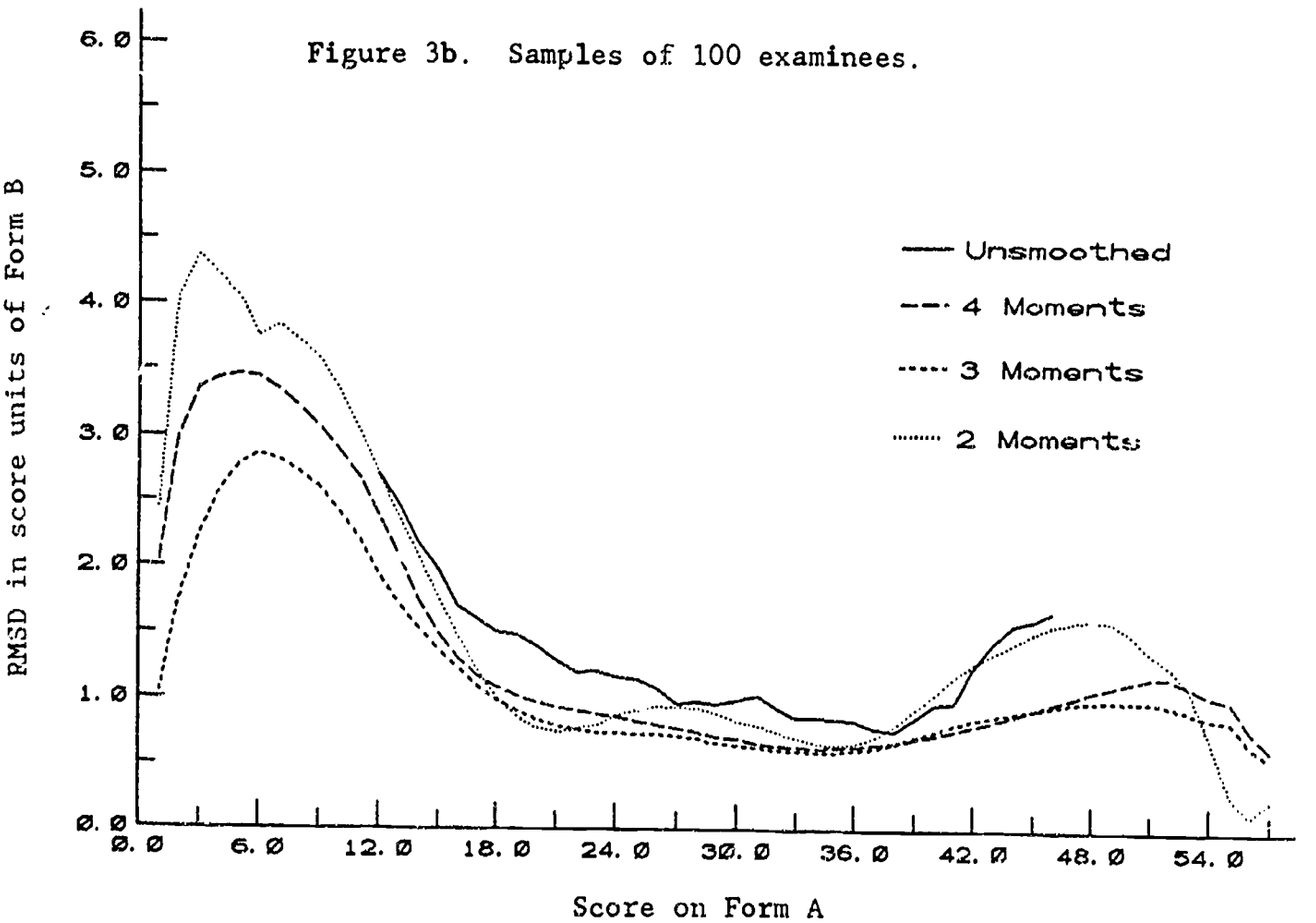


Figure 3 (continued). Conditional root-mean-square difference (RMSD) of sample equating (through common-item anchor) from population equating, over 50 replications of the equating procedure.

Figure 3c. Samples of 50 examinees.

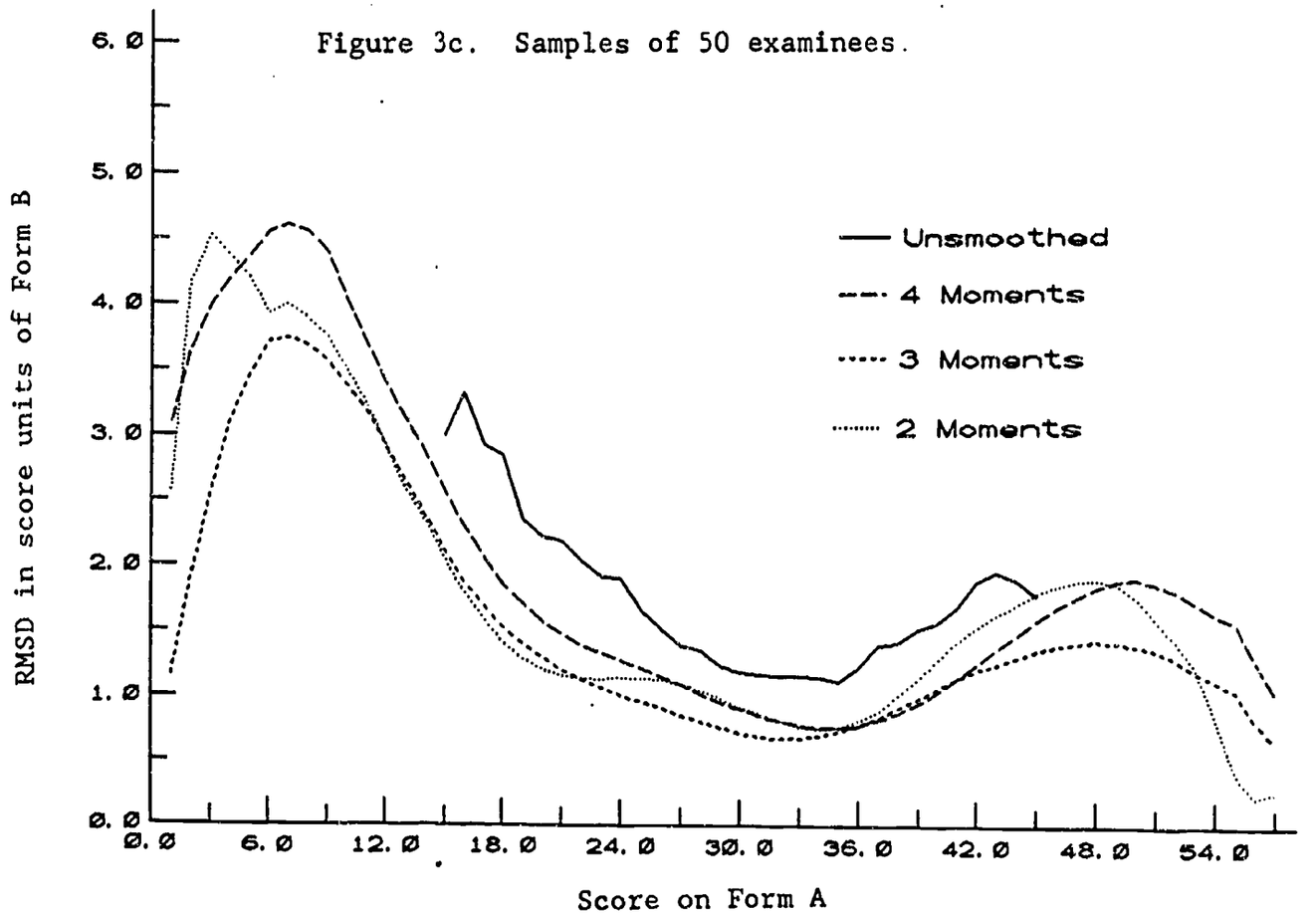



Figure 3d. Samples of 25 examinees.

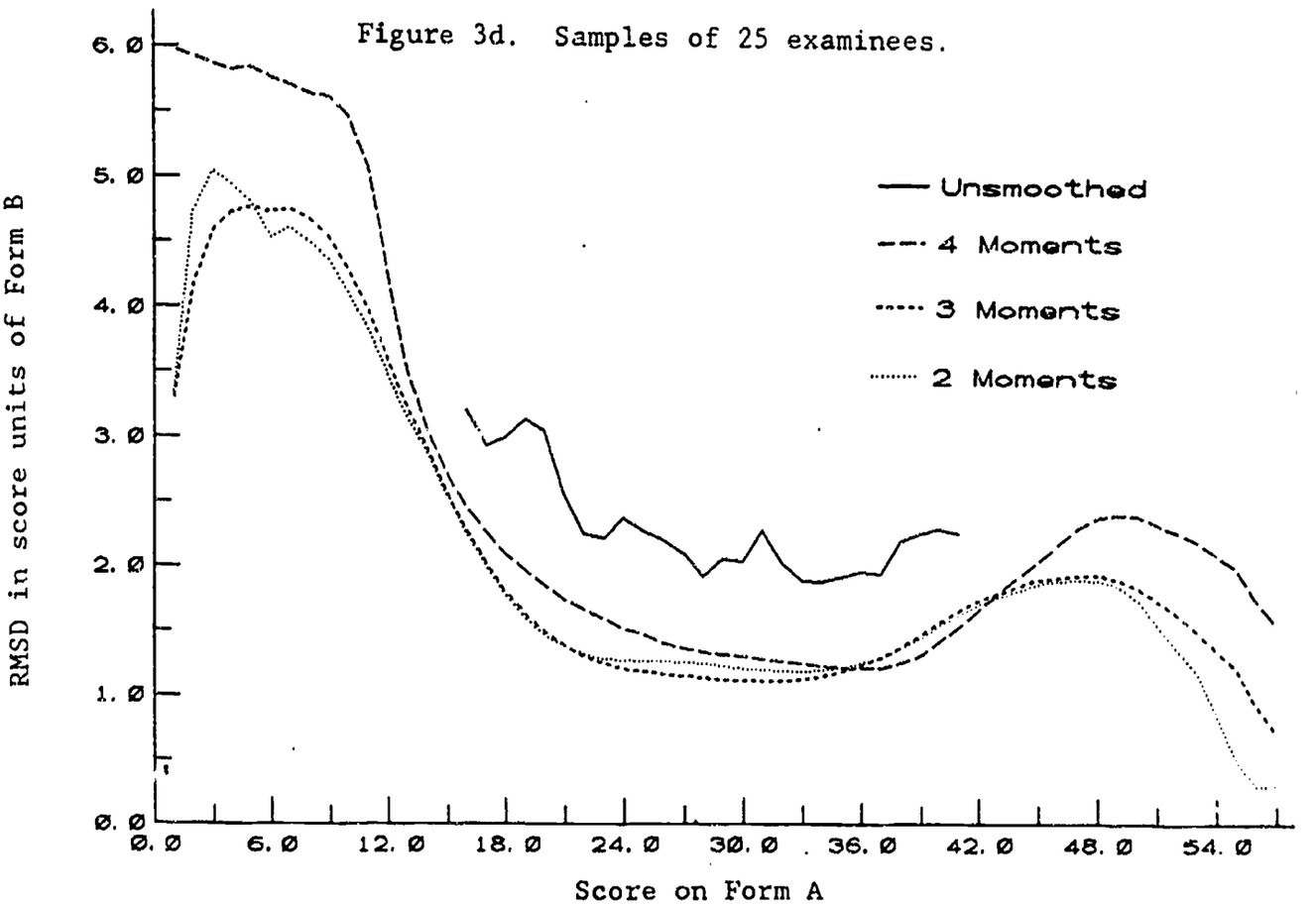


Figure 3 (continued). Conditional root-mean-square difference (RMSD) of sample equating (through common-item anchor) from population equating, over 50 replications of the equating procedure.

Figure 3e. Samples of 200 examinees without smoothing; samples of 100 examinees with 3-moment smoothing.

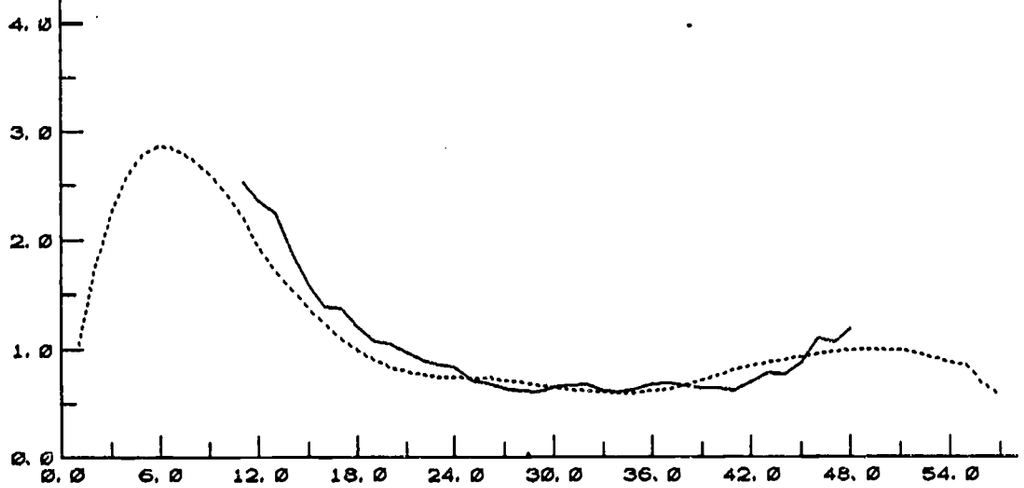



Figure 3f. Samples of 100 examinees without smoothing; samples of 50 examinees with 3-moment smoothing.

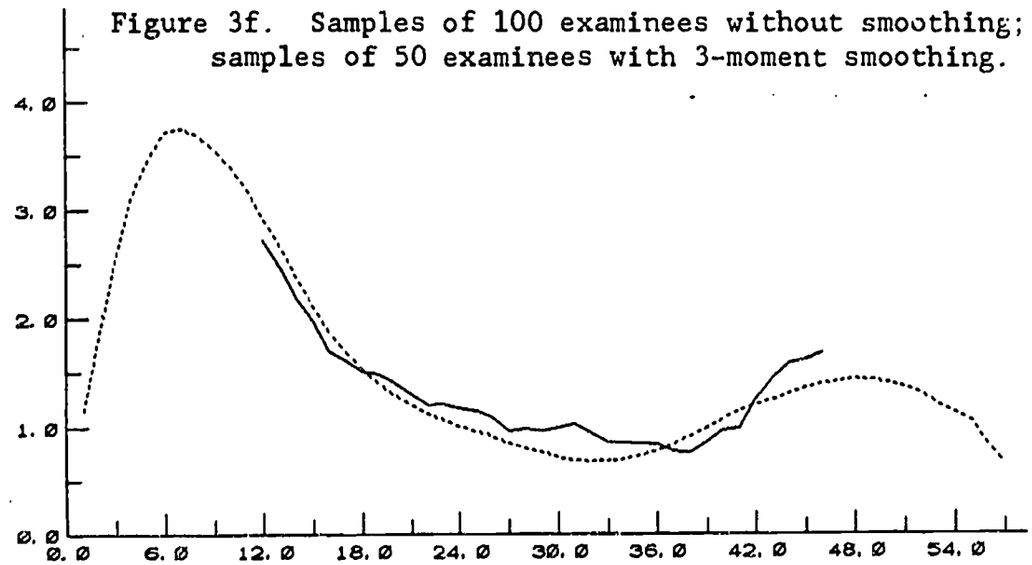
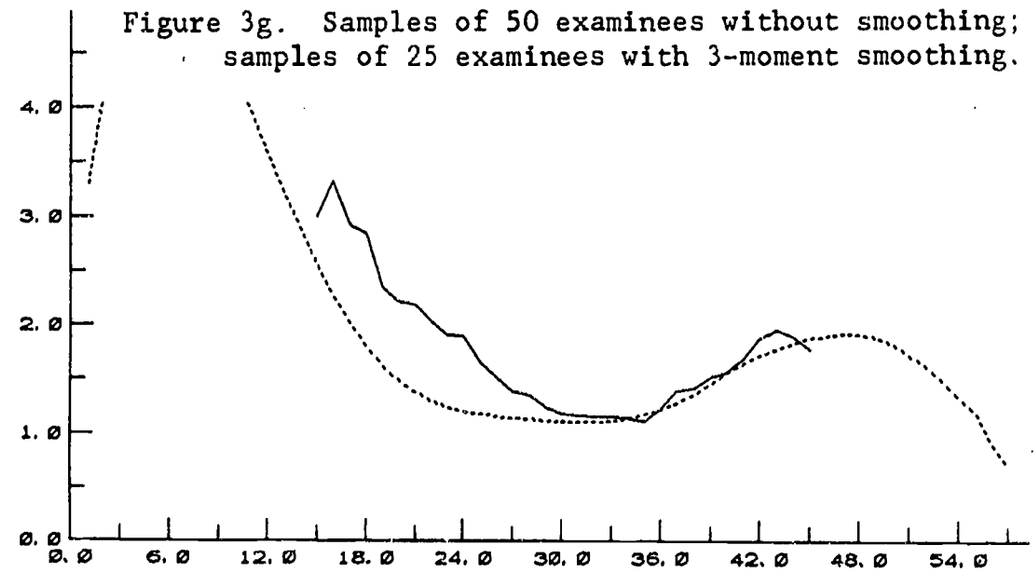
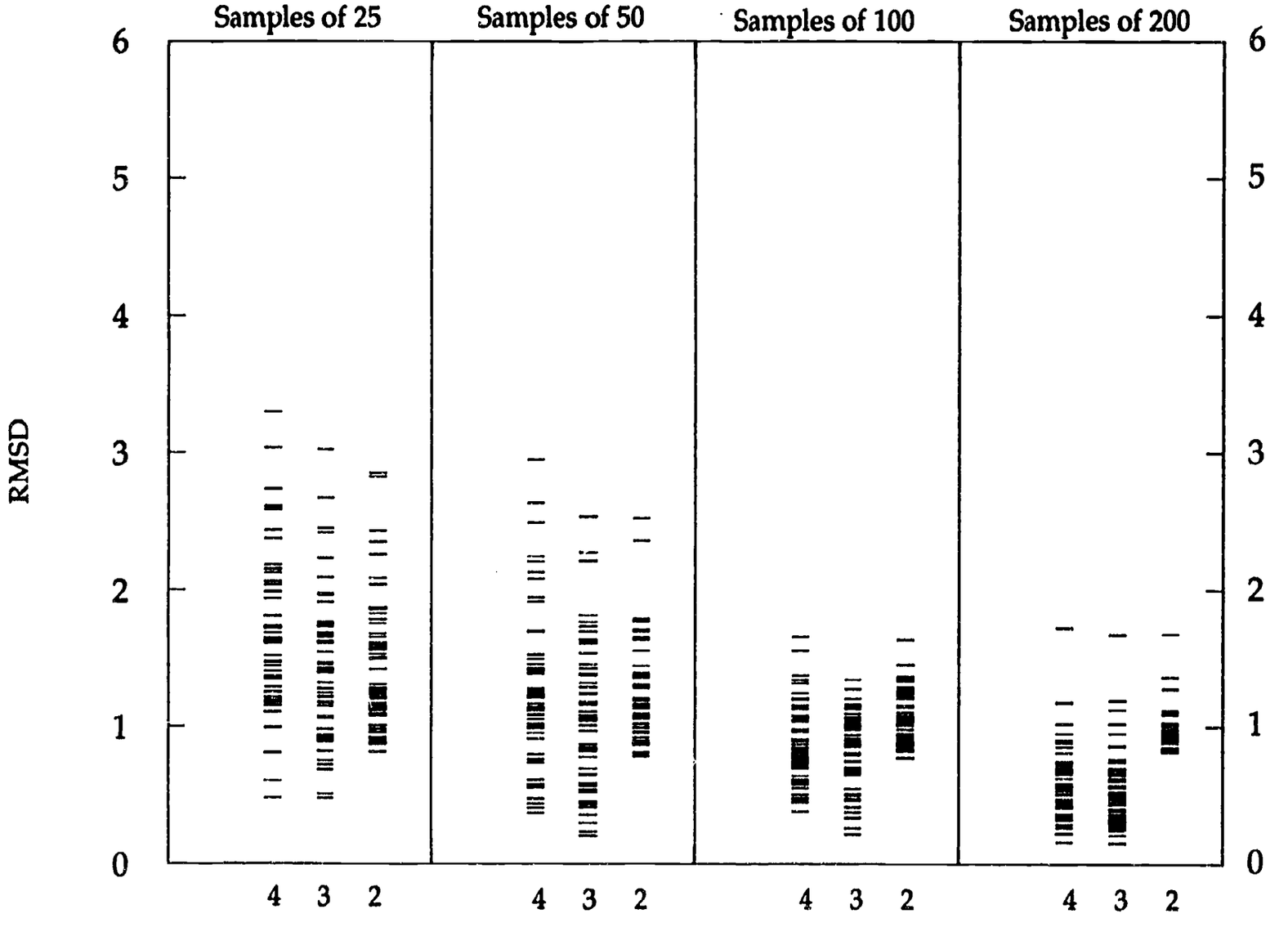



Figure 3g. Samples of 50 examinees without smoothing; samples of 25 examinees with 3-moment smoothing.



# Figure 4.

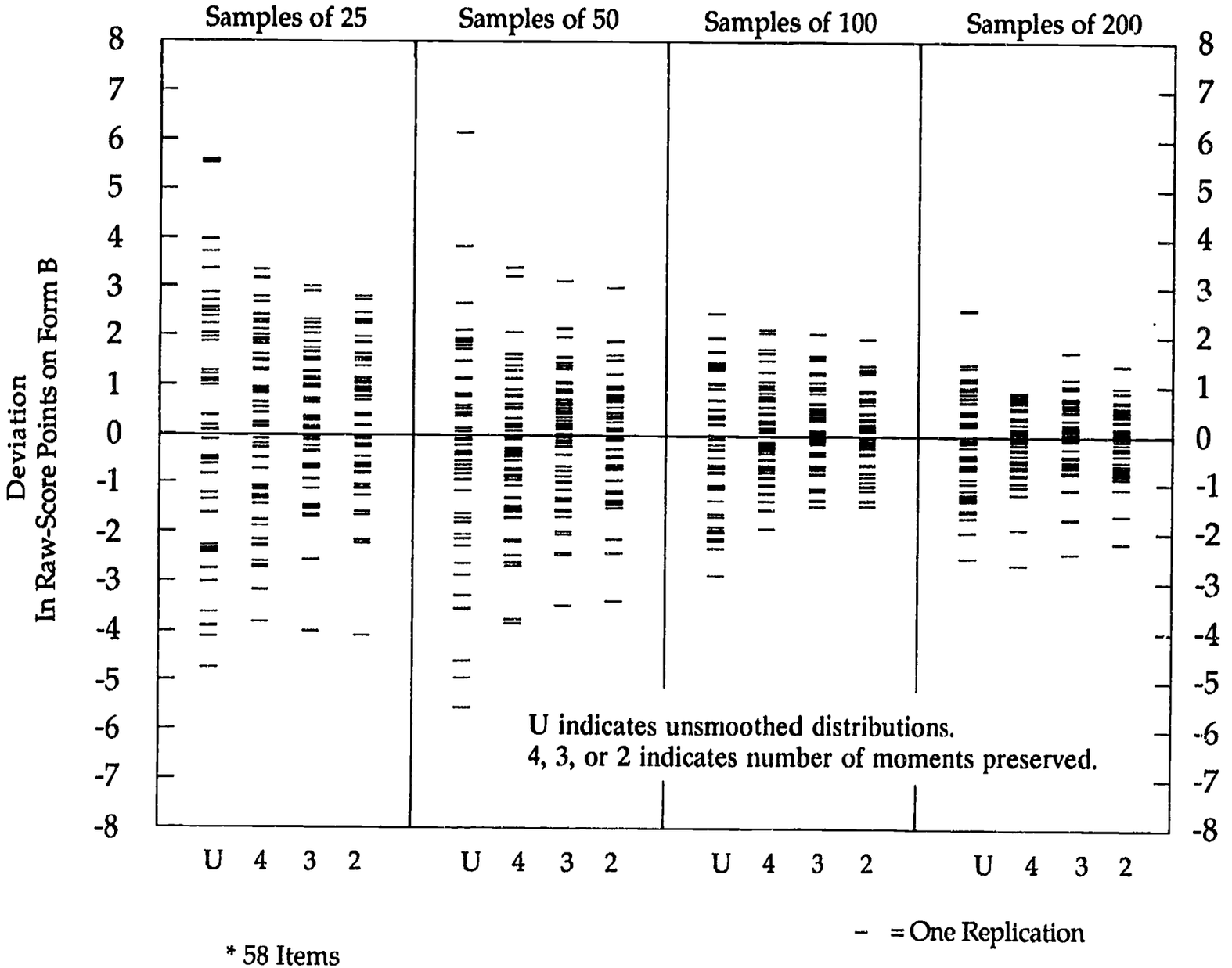
## Overall Weighted RMSD* of Each Sample Equating From the Population Equating




* RMSD over Form A score levels.
Weights are population frequencies.
Units for RMSD are raw score points on Form B.

**Figure 5a.**

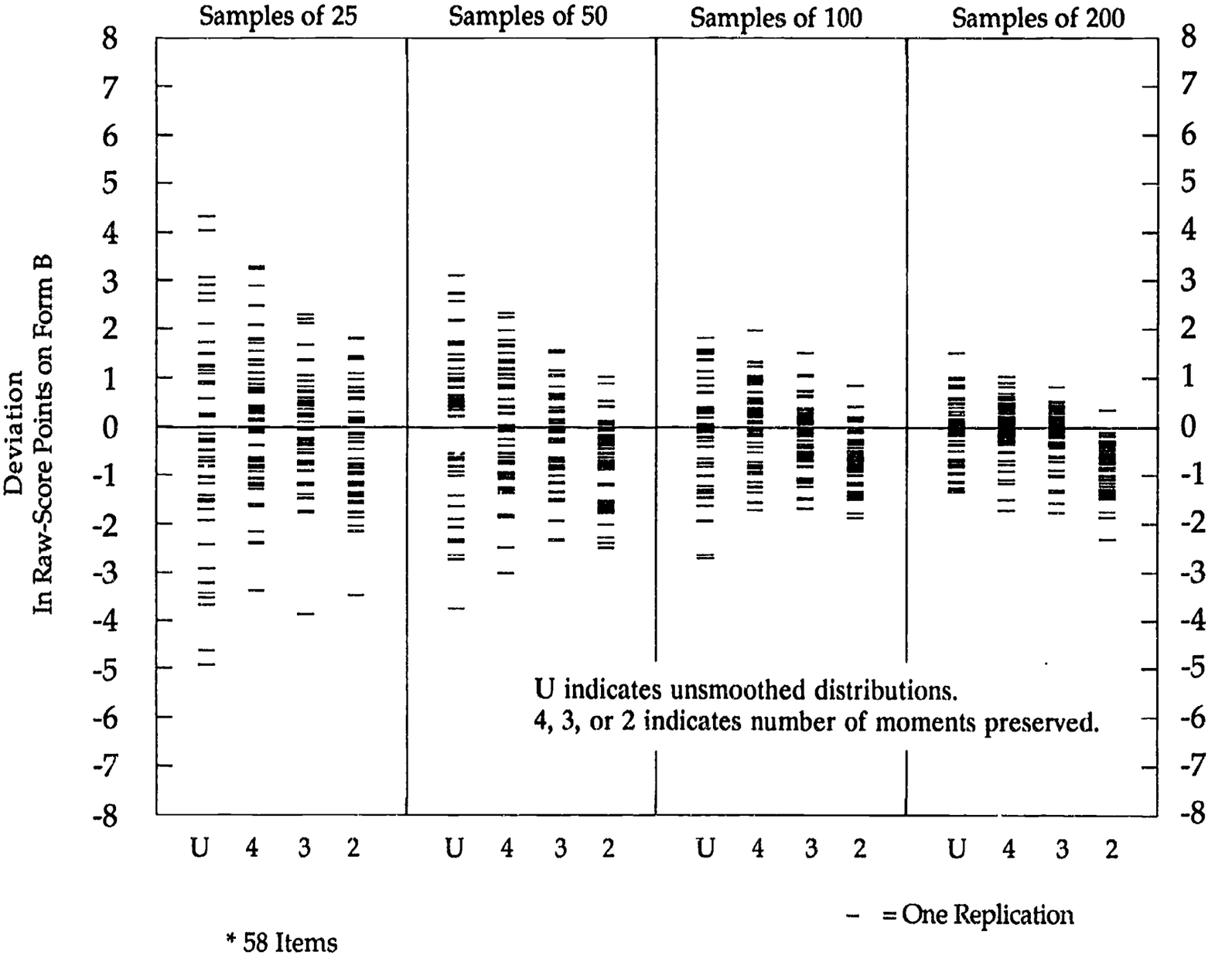
## Deviation of Each Sample (Anchor) Equating From Population Equating At Raw Score 20 on Form A *




* 58 Items

**Figure 5b.**

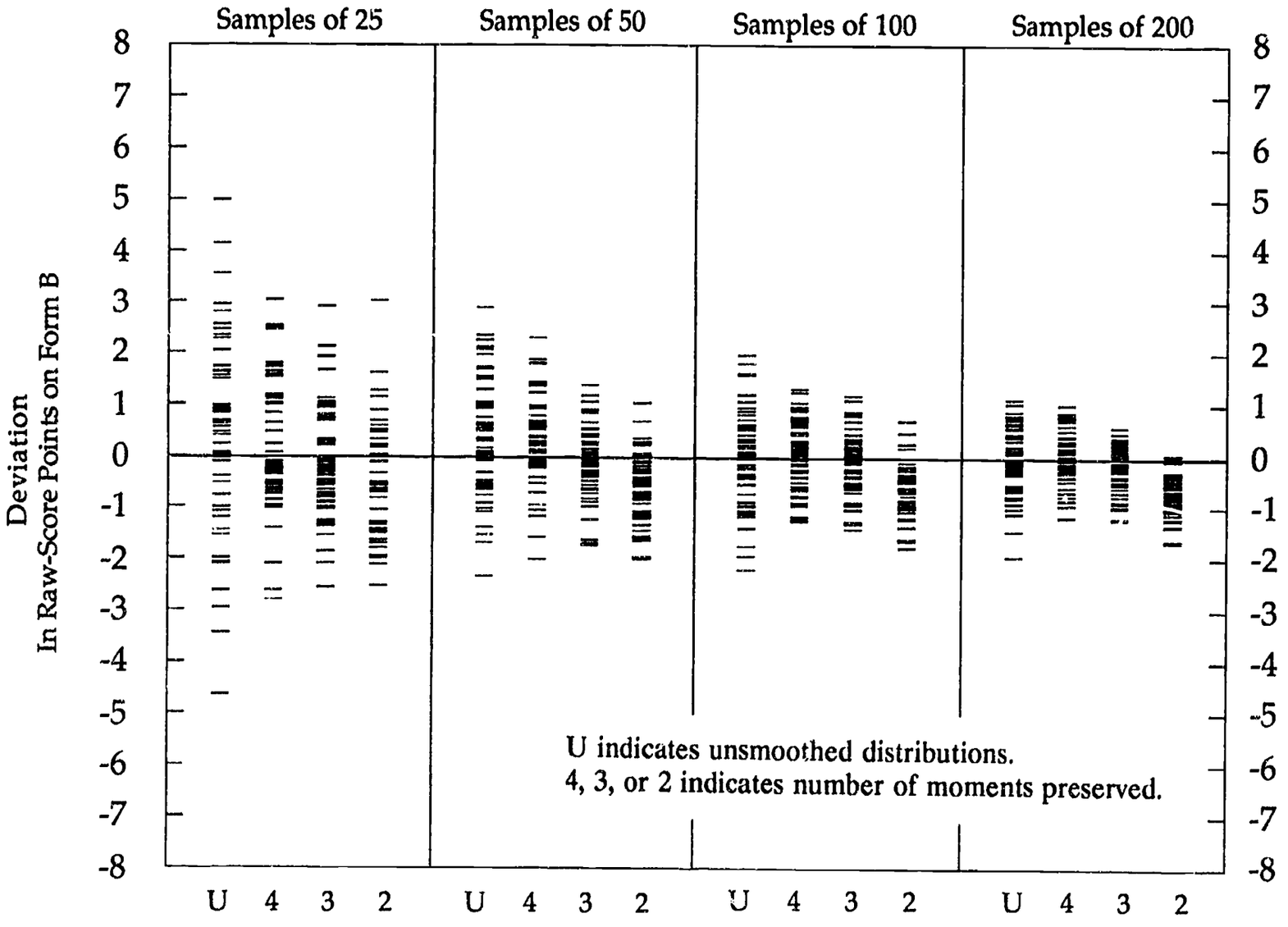
## Deviation of Each Sample (Anchor) Equating From Population Equating At Raw Score 25 on Form A *




* 58 Items

**Figure 5c.**

## Deviation of Each Sample (Anchor) Equating From Population Equating At Raw Score 30 on Form A *

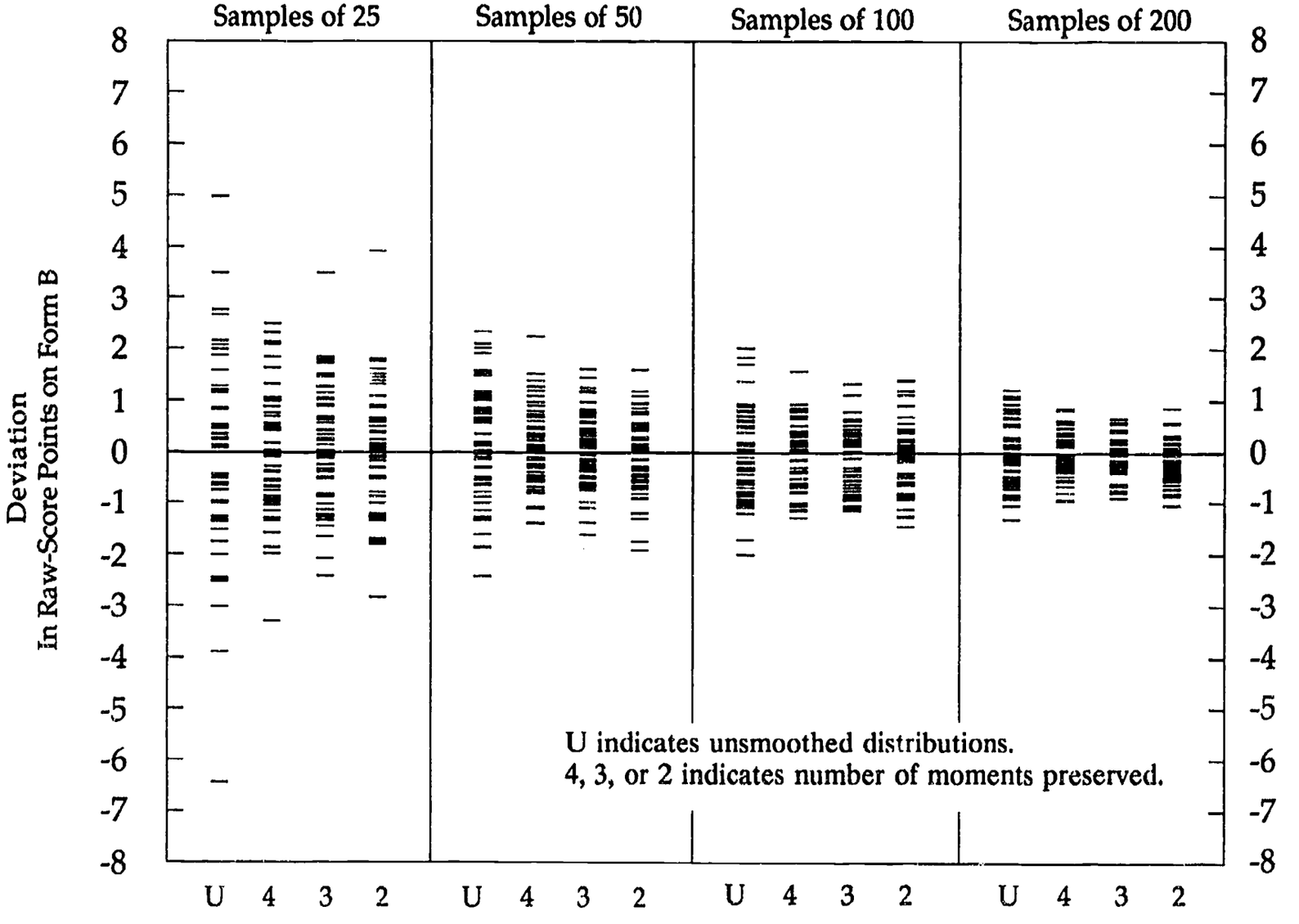



* 58 Items

– = One Replication

**Figure 5d.**

## Deviation of Each Sample (Anchor) Equating From Population Equating At Raw Score 35 on Form A *

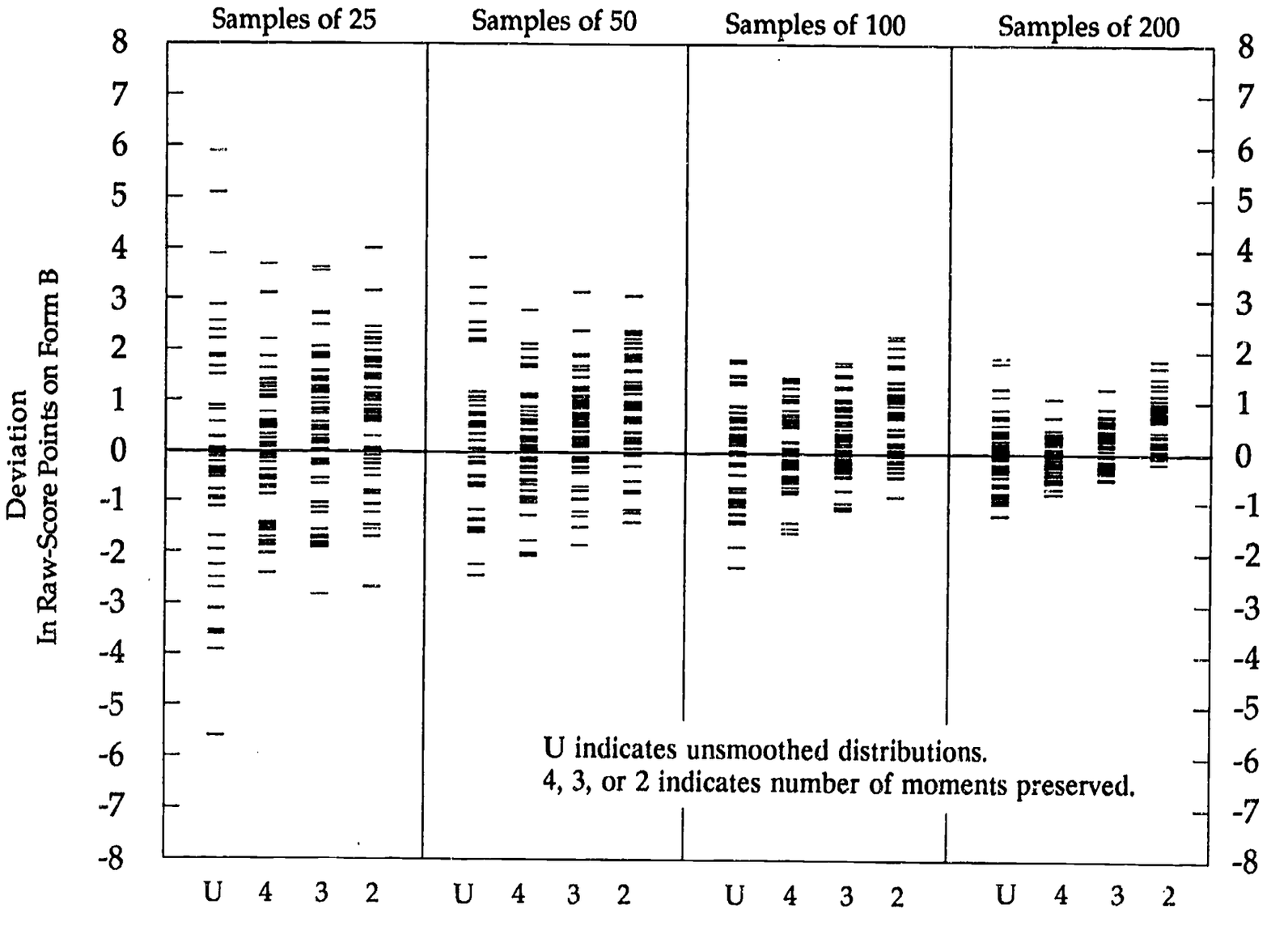



* 58 Items

– = One Replication

**Figure 5e.**

# Deviation of Each Sample (Anchor) Equating From Population Equating At Raw Score 40 on Form A *




* 58 Items

– = One Replication



39

Figure 6. Bias in equating results:
Equated score in samples of 200 examinees (averaged over 50 replications)
minus equated score in population.